MORRISON & FOERSTER LLP

ATTORNEYS AT LAW

3811 VALLEY CENTRE DRIVE, SUITE 500
SAN DIEGO, CALIFORNIA 92130-2332
TELEPHONE (858) 720-5100
TELEFACSIMILE (858) 720-5125

SAN FRANCISCO
LOS ANGELES
DENVER
PALO ALTO
WALNUT CREEK
SACRAMENTO
CENTURY CITY
ORANGE COUNTY
SAN DIEGO

NEW YORK
WASHINGTON, D.C.
NORTHERN VIRGINIA
LONDON
BRUSSELS
BEIJING
HONG KONG
SINGAPORE
TOKYO

RECEIVED
TECH CENTER 1600/2900
03 APR -2 AM 11:52

April 1, 2003

Writer's Direct Contact
(858) 720-5112
KMurashige@mofo.com

SPE Anthony Caputa
U.S. Patent and Trademark Office

Re: U.S. Patent Application No. 09/680,728

Dear Tony,

Here is a color copy of Figure 7 and a draft declaration explaining it.

Could you please let me know if you think this resolves the issue? We were not sure why this figure assumed such importance.

Many thanks,

Kate

Kate H. Murashige

KHM/tmp
Enclosures

sd-140022

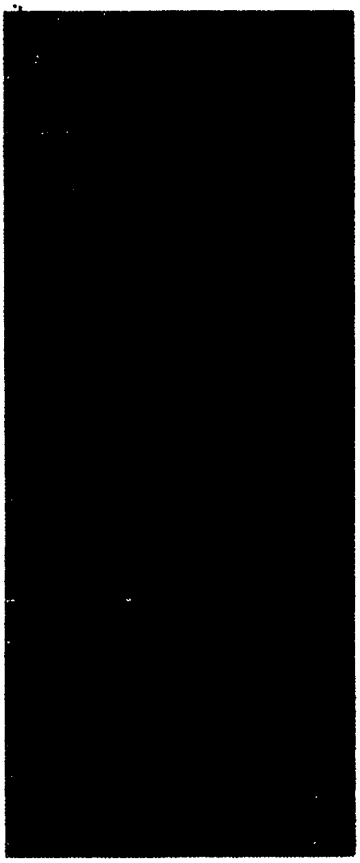


Figure 7. Correlation matrix of 30 protein spots (columns) with mRNA levels as measured by 200 probe-sets on Affymetrix HuFL chips. The correlation coefficients are depicted with colors, bright red being near-perfect correlation (r = 1) and bright green anticorrelation (r = −1). The figure was made using the TreeView software (rana.lbl.gov/EisenSoftware.htm).

Exhibit B

PATENT
Docket No. 511582002400

CERTIFICATE OF MAILING BY "FIRST CLASS MAIL"

I hereby certify that this correspondence is being deposited with the United States Postal Service as first class mail in an envelope addressed to: Assistant Commissioner for Patents, Washington, D.C. 20231, on April ___, 2003.

Tami M. Procopio

RECEIVED
03 APR -2 AM 11:52
TECH CENTER 1600/2900

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In the application of:

Arthur B. RAITANO, *et al.*

Serial No.: 09/680,728

Filing Date: 5 October 2000

For: NOVEL G PROTEIN- COUPLED RECEPTOR UP-REGULATED IN PROSTATE CANCER

Examiner: Minh-Tam B. Davis

Group Art Unit: 1642

DRAFT

DECLARATION OF MARY FARIS

I, Mary Faris, declare as follows:

1. I am currently a Group Leader at Agensys, Inc., the assignee herein. Prior to my employment at Agensys, I was a Senior Scientist at Incyte Genomics. I have a Ph.D. in Immunology and Microbiology from Ohio State University and have held post-doctoral fellowships at the University of Virginia and the University of California at Los Angeles, School of Medicine. While at Incyte, I had considerable experience in expression analysis of cellular mRNA using chips with multiple probes. A copy of my *curriculum vitae* is attached as Exhibit A.

2. I am aware that a question was raised as to the substance of Figure 7 that appeared in an article by Oh, J.M.C., in *Proteomics* (2001) 1:1303-1319. A scanned copy of this Figure, which is in color, is attached as Exhibit B and a copy of the article itself is attached as Exhibit C.

3. The Oh, et al. article discusses a database for use in analysis of protein expression in lung cancers. I agree with the statement made in the Declaration of Pia M. Challita-Eid, that the Oh et al. article would not have merited publication were it not expected that there is a correlation between the existence of mRNA and corresponding protein.

4. The explanation of Figure 7 in the Oh, et al. article is quite brief. It is discussed only at pages 1316-1317 in § 5.2. As stated in § 5.2 and as confirmed in the Figure legend, Figure 7 consists of 30 columns, each representing a protein spot obtained on a 2D gel. For each column, there are 200 entries, one per row, each representing probes for mRNA on Affymetrix' HuFL chips measuring correlation or anti-correlation with the respective protein. The Figure provides 30 discrete sets of data, one set per protein, arranged side-by-side. Thus, the Figure represents selected, detectable proteins on 2D gel and whether a quantitatively similar amount of RNA corresponding to that protein was expressed at one point in time.

5. I cannot identify from the article which 30 proteins are represented; it appears from the explanation in § 5.2 that some of them may be unidentified. Thus, I believe that this Figure is intended to reflect the data presentation concept set forth by Oh et al.

6. I am familiar with the Affymetrix' HuFL chips, and understand that they contain probes for mRNA encoding a wide variety of proteins. Thus, for each individual protein column, at most only some of the probes would even be expected to hybridize with mRNA that actually encoded the protein.

7. As explained in the Figure 7 legend, the level of RNA to protein correlation is represented in color, with red being a near-perfect correlation, green being a negative correlation, and black is not defined. I believe from the results that Figure 7 appears to provide data for two protein "families." The mRNA that are near-perfectly correlated with one family are generally anticorrelated with the other family, as would be expected. This expression pattern correlates to the general quadrants in the Figure.

8. It appears that in all cases there is some mRNA for which a high correlation is found. These data actually support the assertions being made in the present case concerning the qualitative correlation of RNA to protein: In all cases RNA exists that is highly correlated with the existence of the protein (some RNA is simply unrelated to this protein). Thus, each protein perfectly correlated with existence of relevant RNA.

I declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further, that these statements are made with the knowledge that willful, false statements and the like so made are punishable by fine or imprisonment or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

Executed at Santa Monica, California, on ____ April 2003.

Mary Faris

Enclosures:

Exhibit A: *Curriculum Vitae* of Mary Faris

Exhibit B: Figure 7 (color) from an article by Oh, J.M.C., et al. in *Proteomics* 1:1303-1319 (2001).

Exhibit C: A full copy of the Oh, et al., article.

Proteomics 2001, *1*, 1303–1319

Jane M. C. Oh
Franck Brichory
Eric Puravs
Rork Kuick
Chris Wood
Jean Marie Rouillard
John Tra
Sharon Kardia
David Beer
Samir Hanash

University of Michigan
Medical Center,
Ann Arbor, MI, USA

A database of protein expression in lung cancer

We have developed a comprehensive approach to identifying molecular changes in lung cancer that includes both genomic and proteomic analyses. The related effort has produced a large amount of data pertaining to gene expression at the RNA and protein levels. As a result, we have constructed a database that contains protein expression data on lung cancer as well as other relevant data including DNA microarray derived data. A large number of proteins that are expressed in different types of lung cancer have been identified and have been correlated with the expression measures for their corresponding genes at the RNA level. The database is intended to facilitate our effort at developing novel classification schemes for lung cancer and the identification of novel markers for early diagnosis.

Keywords: Lung / Cancer / Database / Microarray

PRO 0131

1 Introduction

There is substantial interest in implementing novel and comprehensive strategies for the molecular analysis of tumors and relevant biological fluids. We have implemented a strategy for the molecular analysis of lung cancer that integrates genomic analysis using genome scanning procedures, transcriptomic analysis using cDNA and oligonucleotide microarrays, and proteomic analysis. For the latter, we have relied to date primarily on 2-D polyacrylamide gels. However the 2-D gel approach is being increasingly complemented with additional analyses using liquid based protein separations and protein microarrays. While on the one hand proteomic analysis complements genomic analysis for a global assessment of gene expression, on the other hand proteomic analysis uniquely contributes an understanding of protein post-translational modifications and the location of protein gene products in subcellular compartments. The scope of our overall molecular analysis study of lung cancer is shown in Fig. 1. Important objectives include the development of novel molecular classification schemes for lung cancer and the identification of novel markers for the early detection of lung cancer.

The large body of proteomic and other data we have collected has necessitated the construction of a database in which basic and derived data is organized. There have been relevant related efforts at databasing of 2-D data by other groups. One such database is the 2DWG Metadatabase of 2-D gel images, which contains 2-D derived data acquired by a combination of review of results as well as submissions by investigators [1]. However, to date there are only three entries found matching the query for human lung images in the 2DWG Web Gel Meta-database web site (http://www-lecb.ncifcrf.gov/2dwgDB). The database we have constructed, in its entirety, is relevant to a variety of cancers. However the focus of this review is the use of the database to achieve our objectives related to the molecular analysis of lung cancer specifically. The goal of the database is to facilitate planned analyses, *i.e.* statistical analysis, as well as post-planned analyses, *i.e.* data mining. The intent is to make the database queryable on a protein – by – protein basis as well as through subgrouping of samples analyzed, in a menu driven fashion. Internet and WWW technologies are used not only to allow investigators to view visual and textual data together, but also to allow investigators in other locations to retrieve archival data using different computer systems.

2 Laboratory information processing system

A long-standing Laboratory information processing system (LIPS) developed by our group [2] has been adapted for our database. LIPS consists of multiple systems and processes. A variety of data is stored in a variety of formats with individualized programs for viewing the data. Typical processes using LIPS include: sample inventory; digitize images; detect and quantify spots; match spots and normalize spot sizes across images, choose spots for MS analysis, enter profiles from MS-Fit web search; transfer data to statistical software or spreadsheets.

Data tend to be complex and dynamic in that their contents are ever changing as information is added, modified or removed. Simple or intensive analyses of 2-D patterns

Correspondence: Dr. S. Hanash, University of Michigan Medical Center, 1150 W. Medical Center Drive, A520 Medical Science Research Building I, Ann Arbor MI 48109–0656, USA
E-mail: shanash@umich.edu
Fax: +1-734-647-8148

Abbreviation: **LTPS**, Laboratory information processing system

© WILEY-VCH Verlag GmbH, 69451 Weinheim, 2001

1615-9853/01/1009–1303 $17.50+.50/0

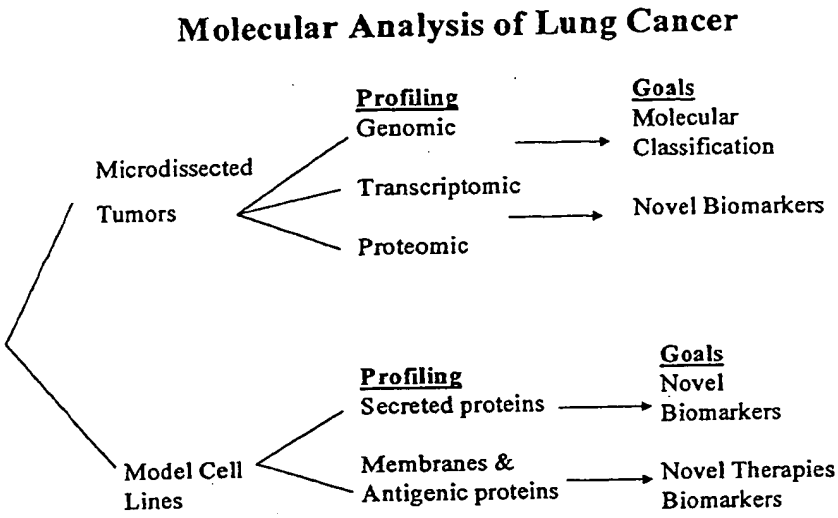


Figure 1. Methods and goals of lung cancer studies.

have produced a large amount of data. Data is both textual (*e.g.*, reports and numbers) and visual (*e.g.*, 1-D and 2-D gel images).

Some types of data generated by LIPS include: 2-D protein gel images (silver, modified silver, blots, ^{35}S labeled gels); genome scans; 1-D gel images; spot information-protein names; gene information from DNA microarrays; MS files and MS-Fit reports (Word documents); figures (Raster files on the Sun and actual photographs); data from protein microarrays; data from liquid chromatography separations.

However, as computer technology has evolved, quantum jumps in improvements in organizing unstructured, scientific data into a structured database have become possible. A major function of our database and its interfaces is to serve as a computer-based tool for capturing the basic quantitative data from 2-D gel images and derived data and findings derived from different studies about proteins detected in 2-D patterns of various tumor types [3]. As a result, investigators are provided with easy access to data as well as a means for intelligent data mining of the existing data. A logical view of the database schema is shown in Fig. 2 and a list of tables and their attributes are shown in Table 1.

The following are important features of the 2-D gel related component of our lung protein database:

(1) All 2-D gel images are placed in hierarchies so that: (a) every study image is matched to one master image, *i.e.* all lung adenocarcinoma tumor images are matched to one master image; (b) every master image is matched to at most one (higher) master image, *i.e.* all masters for different lung tumor types are matched to one tumor master. This allows the database to have an indexing mechanism that can relate a spot to any gel in the hierarchy. The database provides a capability to access the basic and derived data using the following types of queries: (a) given a spot on any gel, find all spots that are matched to it; (b) given a spot on any gel, find all protein identifications made for it, and (c) given a spot on any gel, find all findings/conclusions that are linked to it.

(2) All samples (and thereby gels derived from them) are identified by a list of source characteristics in four major categories: experiment code; cell type code; treatment code; and fraction code. This allows the database to have an identification mechanism that can relate a gel to any source in the hierarchy. The database provides a capability to find all images as follows: (a) given a category, find all images that have the same value of the category; and (b) given any combination of four categories, find all images that satisfy the condition.

(3) All protein spots are identified by a list of characteristics in four major attributes: protein name; p*I* and M_r; accession number; and protein sequence data. A spot may have several findings and there may be many kinds of findings derived from a particular study. If possible the findings are recorded in a consistent way, however this is not always possible due to some characteristics of such findings (*e.g.*, statistical analysis matrices, MS data, and Affymetrix data). As the number of studies has increased, the amount of data produced has increased. Some of the data (*e.g.* mass spectra and Affymetrix (Santa Clara, CA, USA) oligonucleotide chip readouts) is very large, and fills up the hard disks of the computers where it is collected. Such data is generally saved on CD-Rs, and only the most recent data is kept in a computer. It is sometimes easier

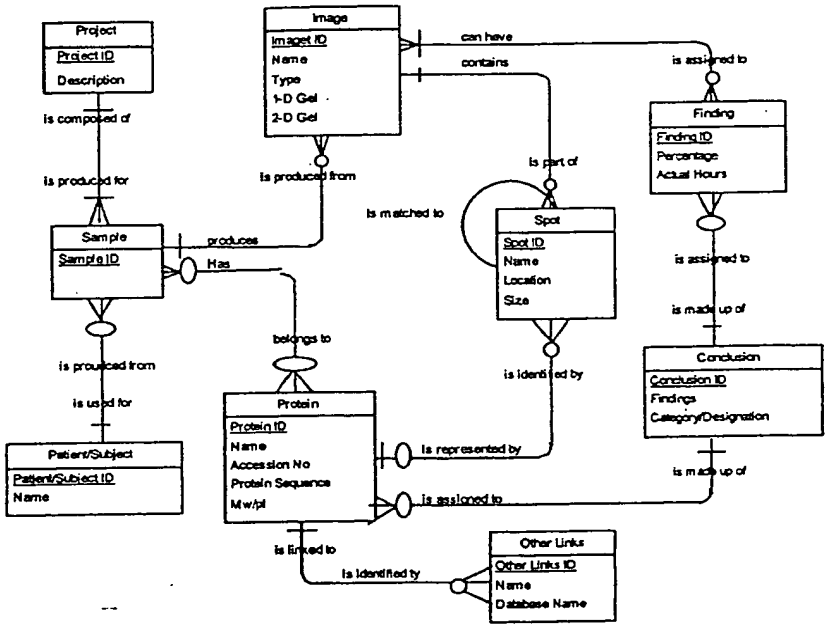


Figure 2. A logical view of database schema.

Table 1. A list of tables and their attributes in the lung protein database

Table name	Unique identifier (Primary Key)	List of attribute types
Project	Project Name	Project Type, Description
Sub Project	Sub Project Name	Date Started, Comment
Subject	Subject ID	Case No, Sex, Birthdate, Comment
Tissue Sample	Tissue Sample ID	Tissue Type, Diagnosis, Date Sample Taken, Date Received, How Received, Source, Comment
DNA Sample	DNA Sample ID	Date Produced, Concentration, Freezer Location, Comment
Gel	Gel ID	Sample ID, Batch ID, Enzyme Combination, Electrophoresis Process, Comment
Image	Image Name	Date Imaged, Exposure Time, Image Type, Image Location, Comment
Spot	Image Name & Spot No	X, Y, Intensity, Spot Type
Match	Match ID	Master Image Name, Master Spot No, Image Name, Spot No
Experiment	Experiment Code	Description
Cell Type	Cell Code	Description
Treatment	Treatment Code	Description
Fraction	Fraction Code	Description
Protein Sample	Sample ID	Experiment Code, Cell Code, Treatment Code, Treatment Date, Fraction Code, Comment, Project Type, Gel ID, Image Name, Image Type, Researcher
Protein	Protein Name	Image Name, Spot No
Other Link	Protein Link ID	Protein Name, Database Name, URL
Findings	Image Name & Spot No	Category, Designation, Finding
Protein Identification	Image Name & Spot No	Accession No, cDNA cloning, Cell Lines, Facility, Date, Genomic Cloning, Glycosylation, M_r, pI, Phosphorylation, Phosphorylation Residues, Related Spot, Sequences, Source of Protein, Name, Structural Modification, Subcellular Localization, Tissue Distribution, Type of Membrane, Type of Sequencing

to post individual files on the web. Individual web pages have been created with textual and visual data that are difficult to relate in a table. This allows investigators an opportunity to analyze 2-D gel and other images containing spots that have not been detected or identified and to compare data across studies. In addition this is used to link our data to other biological knowledge repositories such as GenBank, PIR International, and SWISS-PROT.

3 Contents of the lung cancer protein database

A large number of studies involving lung cancer have been independently performed in the laboratory. At the protein level, these studies have resulted in 1349 images, over 1000 of which are images of 2-D gels for which information has been recorded in the lung protein database. This number represents a fraction of the 30 682 2-D gels produced by our group for different studies, which include studies of other cancer types encompassing head and neck, esophagus, liver, colon, pancreas, ovary, breast, prostate, brain, leukemias and childhood tumors. A list of protein gel images related to lung studies is shown in Table 2. While lung adenocarcinomas represent a major portion of the database, other lung tumor types including squamous cell carcinomas and small cell lung cancers are represented, as are control lung tissues. Other 2-D patterns were produced from studies of cell lines that have been manipulated by transfection or by treatment with specific agents, as well as patterns produced after different cell fractionation schemes. Substantial emphasis is currently being placed on the comprehensive profiling of lung cancer derived surface membrane proteins.

Table 2. A high-level categorization of lung protein 2-D images by sample type

Lung Sample Types	
Cell Lines	421
Cystic Fibrosis	44
Tumor	635
Normal	170
Plasma	61
Other	18
Total	1349

Mass spectrometry and/or *N*-terminal sequencing of protein spots from 2-D gels of lung tumor samples or cell lines have led to the identification of a large number of proteins expressed in lung cancer. Also, most identifications made for proteins from a sample type can often be confidently transferred to matching protein spots on master images from lung studies. Table 3 and Fig. 3 exhibit some of the progress we have made in identifying proteins in 2-D gels of lung samples.

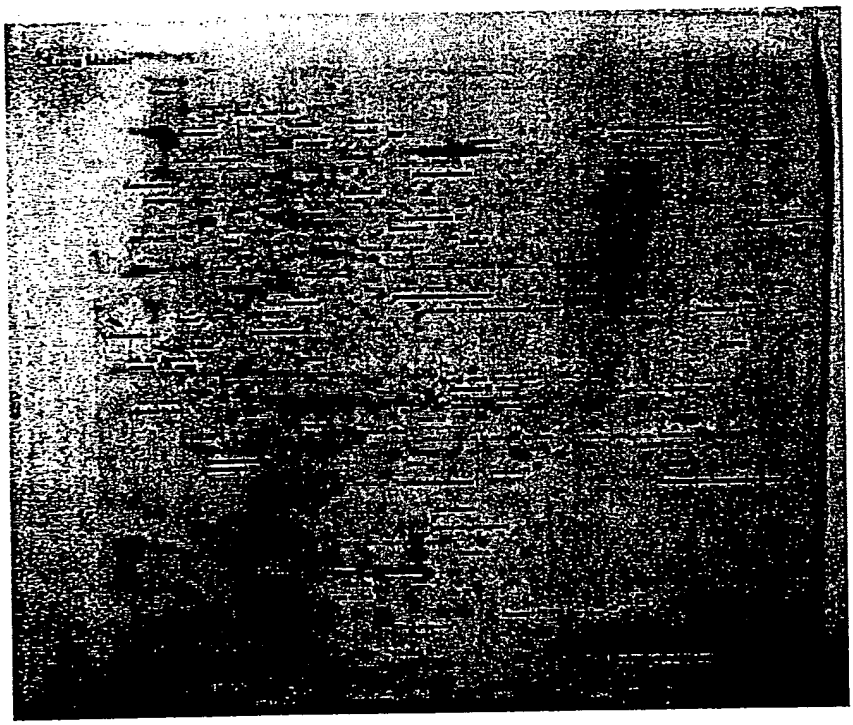


Figure 3. Small cell lung tumor master containing identified proteins.

Table 3. A list of identified proteins

ID Source	Name	Spot #	NCBI Accession Number	GenBank Number	*pI*	M_r	Official gene
L95	(spot 1496L) possibly pacreatitis-associated protein	1496					
L M	14_3_3_sigma	577	398953	P31947	4.361	30.052	SFN
L M	14_3_3_ZetaDelta	615	112695	P29312	4.569	29.101	YWHAZ
L95	14–3–3n	1279	437363	AAA35483			YWHAH
DMS 79	6PF-2-K/FRU-2,6-P2ASE Liver isozymer	24	2507178	P16118			PFKFB1
	ADP-ribosylation factor 1	928	4502201	NP_001649	6.31	20.697	ARF1
	Albumin	319					ALB
	Albumin	800					ALB
L M	Albumin				5.957	70.244	ALB
L M	Aldehyde Dehydrogenase	207	4502031	NP_000680	6.811	56.966	ALDH1A1
L M	AldoKeto Reductase	543	3493209	AAC36469	7.812	32.379	AKR1B10
SKMES	Alkaline Phosphatase, Placental type 1 precursor	14	130737	P05187	5.86	57.954	ALPP
	Albumin	666					ALB
	Albumin	693					ALB
L M	α-Enolase	332	4503571	NP_001419	7.742	45.407	ENO1
L95	α-helical protein	268	8272482	AAFZ4221			HCR
L95	α-helix coiled-coil rod homolugue	268	5360901	BAA82158			
L M	α Tublin	172	5174477	NP_006073	5.099	52.848	
L M	Amyloid B4A	802			4.796	17.194	
A549	Annexinl	460	113944	P04083	6.73	39.264	ANXA1
L M	Annexin V	522	4502107	NP_001145	4.83	33.326	ANXA5
L M	ApoAI	685			5.124	25.4	
L95	Apoprotein, pulmonary surfactant	1278	71967	LNHUPS			
L M	ARF1	795			6.33	18.909	ARF1
SKMES	β-Galactoside soluble lectin	96	227920	1713410A	5.34	14.584	LGALS1
LM	β-Actin	349	113270	P02570	5.29	41.7	ACTB
DMS 79	β-spectrin	61	29497	X59511			SPTB
LM	β Tubulin	229	4507729	NP_001060	4.75	49.8	TUBB
DMS 79	Calmodulin dependant phosphodiesterase	22	11995077	AB038211			
L M	Calreticulin	104	4757900	NP_004335	3.668	57.29	CALR
L M	Calreticulin32	469			3.442	48.772	
A549	CGI–46 protein	36	4929561	AAD34041	6.25	49.296	
A549	Chaperonin–like protein	149	4502643	NP_001753	7.034	60.547	CCT6A
L95	Clathrin light chain A	1338	4502899	NP_001824			CLTA
	Collagen, type XV, α 1	789					COL15A1
DMS 79	Complexin II	85	1362772	E57233			CPLX2
L M	Cellular retinoic acid-binding protein 2	856			5.415	11.858	CRABP2
L M	Cellular retinol-binding protein 1, CRPB1	855	4506451	NP_002890	4.667	10.297	RBP1
	Creatine kinase, brain	439	180570	AAC31758	5.34	42.618	CKB
L M	Cytochrome C bxydase VA	872			4.568	9.2	
A549	Cytokeratin 8	321	1673575	U76549			KRT8

Table 3. Continued

ID Source	Name	Spot #	NCBI Accession Number	GenBank Number	pI	M_r	Official gene
A549	Cytokeratin 8	446	2506774	PO5787	5.52	53.674	KRT8
A549	Cytokeratin 8	439	2506774	P05787	5.52	53.674	KRT8
L M	Cytokeratin 15, keratin 15	289	4504915	NP_002266	4.153	49.261	KRT15
A549	Dihydrolipoamide dehydrogenase, mitochondrial precursor	759	118674	P09622			
L M	DJ1	811	6005749	NP_009193	6.44	21.015	DJ1
L M	DJ1_MER5	700			6.263	24.001	
DMS 79	dj475N16.1 (CTG4A)	57	6969163	CAB75301			
L M	DUTPhase	769			5.719	20.136	
L95	E2 ubiquitin-conjugating enzyme	1445	4885417	AB022435			HIP2
L M	EIF4d	718			5.104	22.961	
L M	EIF5A	839			4.599	10.957	
	Enhancer of rudimentary (Drosophila) homolog	902					ERH
	Enolase 2 (γ, neuronal)	295	119347	P09104	4.94	47.286	ENO2
L M	ENPL_HSP100	18			4.945	78.717	
A549	F1FO-type ATP synthase subunit d	1519	5453559	NP_0063475	5.21	18.491	ATP5JD
DMS 79	G1/S specific cyclin E1	31	3041657	P24864			CCNE1
L M	G3PD	540			7.457	31.772	
L M	γ-Actin	348	113278	P02571	5.146	42.315	ACTG1
L M	Glyoxalase1	650	417246	Q04760	4.833	25.572	GLO1
FMD 79	Granulocyte-macrophage colony-stimulating factor precursor	86	117561	PO4141			CSF2
L M	GRP75	87			5.9341	73.124	
L M	GRP78	79			5.187	68.109	
L M	GSTpi	690	726098	AAC13869	5.5	25.4	GSTP1
	Heat shock 27 kD protein 1	626	123571	PO4792	7.83	22.327	HSPB1
	Heat shock 27 kD protein 1	631	123571	PO4792	7.83	22.327	HSPB1
A549	Heterogeneous nuclear ribonucleoprotein H	457	5031753	NP_005511			HNRPH1
A549	HLA-B71 or HLB-B71 variant	818	511776	U11269	5.55	36.558	
L M	HSC70_HSP73	120			5.893	72.429	
L M	HSP90	46			5.276	76.096	
L95	HSPC089	1036	6841118	AAF28912			
L95	HSPC321	1547	6841292	AAF28999			
L95	HSPC321	1548	6841292	AAF28999			
A549	HuCha 60 SP 60	181	4504521	NP_002147	5.7	61	HSPD1
L95	Huntingtin associated protein	1595	1708113	P54255			HAP1
L95	Huntingtin associated protein	1548	1708113	P54255			HAP1
	Internexin neuronal intermediate filament protein, alpha	183	6225015	Q16352	5.48	54.908	INA
	Keratin 17	934	4557701	NP_00413	4.97	48.106	KRT17
DMS 79	KIAA1610 protein	26	10047295	AB046830			
L M	LamR	340			4.549	44.03	

Table 3. Continued

ID Source	Name	Spot #	NCBI Accession Number	GenBank Number	*pI*	M_r	Official gene
	Lectin, galactoside-binding, soluble, 1 (galectin 1)	873	227920	1713410A	5.34	14.584	LGALS1
L M	Lipocortin	460	113944	PO4083	6.73	39.264	ANXA1
A549	L-Lactate Dehydrogenase H chain	906	126041	PO7195			LDHB
A549	L-lactate dehydrogenase H chain (LDH-B)	906	4557032	NP_002291			LDMB
L M	LaminB	924			5.787	69.625	
	Lymphocyte cytosolic protein 1 (L-plastin)	924	4504965	NP_002289	5.20	70.290	LCP1
L95	Macropain subunit zeta	1338	4506187	NP_002289			PSMAS
DMS 79	MHC class 1 histocompatability antigen protein	33	1236790	U06487			
DMS 79	Multicalaytic endopeptidase comples chain C2, long splice from	74	346314	JC1445	6.51	30.239	
L M	MyosinLightCahin3	815			4.11	15.172	
A549	Nm23, NDPKA	1456	127981	P15531	5.809	19.216	
	Non metastatic cells 1, protein (NM23A)	793	4557797	NP_000260	5.83	17.148	NME1
L M	Op 18, leukemia-associated phosphoprotein p18 (stahmin)	809	5031851	NP_005554	5.783	17.164	LAP18
L M	Op 18a	807	5031851	NP_005554	4.962	13.655	LAP18
L M	Op 18m	808	5031851	NP_005554	5.302	14.857	LAP18
L M	Phosphoglycerate MutB	639			7.083	27.227	
L M	Phospholipase C	248			5.7	56.5	
L M	PIMT	662			6.211	25.804	
L95	Pinch-2 protein	1695	9800509	AAF99328			
L95	Pinch-2 protein	1825	9800509	AAF99328			
L95	Possibly activin type II receptor precursor; DNA polymerase epsilon subunit B; or ITF-1 DNA binding protein	627					
L95	Possibly BTF2p44	1496					
A549	Possibly carbonci anhydrase III or UCH-L1; PGP 9.5	1242					
A549	Possibly δ-3,5 δ-2,4-Dienoly-CoA isomerase precusor	2138					
A427	Possibly G1 to S phase transition protein; serine-theonine phosphatease protein; or phosphatase 5 protein	321					
L95	Possibly GCF2 fusion protein or Bamacan homolog	320					
L95	Possibly glycosyltransferase	1519					
L95	Possibly HLA DQ	1271					

Table 3. Continued

ID Source	Name	Spot #	NCBI Accession Number	GenBank Number	*pI*	M_r	Official gene
A549	Possibly hydroxyacylglutathione hydrolase or B-lymphocyte Antigen CD20	1080					
L95	Possibly microtubule-based motor protein	1438					
L95	Possibly putative novel protein similar to HPS	1427					
L95	Possibly Spi-B; unnamed protein product (AK001844); or protein kinase (γ15801)	1187					
L95	Possibly T-complex protein	630					
A549	Possibly U 1 small nuclear ribonuclear protein Á	1148					
L95	Possibly unnamed protein product (AK000369) or syntaxin	1064					
L95	Possibly unnamed protein product or Pro0282p protein	1351					
	procollagen-proline, 2-oxoglutarate 4-dioxgenase (proline 4-hydroxylase), beta polypeptide (protein disulfide isomerase; thyroid homone bindung protein p55)	110	2507460	PO7237	4.76	57.116	P4HB
	proliferating cell nuclear antigen	515	129697	P17070	4.4	37.5	PCNA
	Protein phosphatase 2 (formerly 2A), regulatory subunit A (PR 65), β-isoform	104	5915686	P30154	4.84	66.202	PPP2R1B
L M	Protein H precursor	40			3.714	62.182	
L M	Protein kinase C inhibitor 1	882	4885413	NP_005331	7.714	11.521	H1NT
L95	Pulmonary surfactant apoprotein precusor	1278	190565	AAA36510			SFTPA1
L95	Pulmonary surfactant-associated protein	1278	131412	PO7714			SFTPA1
L M	R33729_1	848	3355455	AAC27824	7.508	13.163	
	Retinol-binding protein 1, cellular	855	4506451	NP_002890	4.99	15.850	RBP1
L M	RoSS_A_Antigen	69			3.215	47.903	
	S100 calcium-binding protein A11 (calgizzarin)	906					S100A11
	S100 calcium-binding protein A8 (calgranulin A)	910	115442	PO5109	6.51	10.834	S100A8
	S100 calcium-binding protein A9 (calgranulin B)	931	6094219	P50117	6.37	13.291	S100A9
DMS 79	Serine/threonine protein phosphatase 2A, 65kDa regulatory Subunit A, β isoform	14	5915686	P30154	4.84	66.202	PPP2R1B
	SET translocation (myeloid leukemia-associated)	376	1711383	Q01105	4.12	32.103	SET

Table 3. Continued

ID Source	Name	Spot #	NCBI Accession Number	GenBank Number	*pI*	M_r	Official gene
	Small glutamine-rich tetraricopeptide repeat (TPR)-containing	476	8134666	O43765	4.81	34.063	SGT
	Stratifin	577	398953	P31947	4.68	27.774	SFN
L M	Superoxidedism CuZn	792	134611	P00441	5.6	17.3	SOD1
L M	Superoxide DismMN, superoxide dismutase 2, mitochondrial	737	134665	PO4179	7.887	20.78	SOD2
L M	TCP 1 β subunit	202			5.89	59.841	
L M	TCTP (translationally-controlled tumor protein 1)	680	4507669	NP_003286	4.688	25.143	TPT1
L M	Thioredoxin	896			4.689	9.207	
L M	Tplastin HSP 70	125			5.862	68.909	
L M	Transthyretin	842			5.693	14.714	
A549	Triosephosphate isomerase	672	136060	P00938	7.2	25.5	TPI1
L95	Tropomyosin, cytoskeletal type, tropomyosin 5	550	136096	P12324	4.5	31.9	
LM	Tropomyosin 4	548	13274400	AAK17926	4.377	32.733	TPM4
L95	Troponin T	866	408217	AAB27731			
L95	Troponin T	778	408217	AAB27731			
	Tublin, β polypeptide	229	4507729	NP_001060	4.78	49.907	TUBB
DMS 79	Tumor associated hydroquinone (NADH) oxidase tNOS	34	6644167	AF207881			
	tyrosine 3-monooxygenase/ tryptophan 5-monooxygenase activation protein, epsilon polypeptide	576	1168198	P4266	4.63	29.174	YWHAE
	tyrosine 3-monooxygenase/ tryptophan 5-monooxygenase activation protein, zeta polypeptide	615	112695	P29312	4.73	26,645	YWHAZ
	Tyrosine 3-monooxygenase/ tryptophan 5-monooxygenase activation protein, theta polypeptide	579	112690	P27348	4.68	27.764	YWHAQ
A549	Ubiquitin carboxyl-terminal esterase L1 (ubiquitin thiolesterase), UCH-L1; PGP 9.5, GST mu	656	136681	PO9936	5.283	27.745	UCHL1
L95	Unnamed protein product	1270	7023092	BAA91833			
A549	Urokinase plasminogen activator	842	487123	S39495	6.01	31.263	
L M	Vid1	293			4.712	47.485	
L M	Vid2	294			4.614	46.369	
L M	Vid4	337			4.464	45.322	
	Vimentin	294					VIM
A427	Vimentin	606	4507894	NM_003380			VIM
A549	Vimentin	505	418249	PO8670			VIM
A549	Vimentin	47	340234	M25246			VIM

In addition to 2-D gel analysis, most lung adenocarcinomas are examined at the genomic level using restriction landmark genome scanning, and by mutation analysis for a small number of genes. Transcriptomic analysis is done primarily using oligonucleotide microarrays, as part of our efforts to derive a molecular based classification of lung adenocarcinomas that is more predictive of clinical behavior for this group of tumors than current classification schemes. We also have similar molecular analyses of control lung tissue obtained from multiple sources including adjacent lung tissue from lung cancer patients as well as tissues obtained from non-cancer resected lung.

Only a fraction of the information in the 2-D patterns has been linked across all studies and analyses. The lung protein database contains the basic descriptive data of various samples analyzed, the images of the 2-D patterns that resulted from these samples, the quantitative spot data and information about which spots have been matched to each other, and conclusions or findings about spots. The database is intended to allow not only the retrieval of existing data, but also to mine new information and knowledge about protein expression in lung cells. Data mining activities consist, for example, of reviewing previous studies and finding out which 2-D gel patterns and protein spots are interesting for post-planned analysis and new discoveries. Such discoveries derive from: (1) identification of proteins that exhibit interesting expression profiles in 2-D patterns that have been regrouped from different experiments and studies; (2) expanded statistical analyses that cover protein expression patterns involving large numbers of experiments and images; (3) relating our data involving proteins to outside information; and (4) relating proteomic data to genomic data.

4 Use of the database for post planned analysis

4.1 Virtual matching

Interactive software packages are used to automatically detect and quantify spots and to match spots between different protein patterns, with visual editing to correct any errors in computer based matching. The spot match program has created indices that allow investigators to quickly navigate through many gels and easily compare spots on images from many different experiments and studies, discover proteins of interest, and access and view relevant data. Here the term "match" is used as a logical "transitive" relation, which means if spot A is matched to spot B and spot B is matched to spot C then the spots A and C are considered matched. The lung protein database contains data on proteins detected on various 2-D gels. Since all gels derived from whole cell or tissue lysates in the lung protein database are tied into a single hierarchy, protein identification data recorded for a spot is used to derive protein data for its matched spots using an advanced query capability of the database. This is known as "virtual matching" or "virtual protein identification", which allows investigators to access and view all matched images and the corresponding information from the lung protein database. With a click on a spot, one gets the result shown in Fig. 4. The virtual protein identification feature does not provide a 100% level of certainty of protein identification, but it makes possible the display of spots of interest. A combination of automated recognition and manual editing generally yields an accurate record of protein information in the database for previously unknown proteins. With this approach, the lung protein database will evolve and mature to include all correct data for further analysis and data mining.

4.2 Integrating protein spot data with MS data

As interest in proteomic analysis grows, a number of very large public databases are available to access protein data *via* the internet. Public databases offer a sophisticated text search and keyword search, which links any entered keyword to all protein information associated with that keyword, to ensure easy access to all relevant data. Protein identification using MALDI-MS relies on database searches and usually has three components: (1) peak detection which allows automatic determination of peptide masses; (2) search in protein sequence databases (SWISS-PROT and/or GenBank) for protein entries that match the masses; and (3) certainty calculation which determines the quality of the match for each protein in the list [4]. An example of such a software tool is the PepFrag for searching protein and DNA sequence databases that can use different types of mass spectrometric information [5]. Fenyo [6] described methods and software tools in proteomics for identifying and characterizing proteins, which emphasizes MS combined with database searching. Proteolytic peptide mapping and genome database searching provide an automated means for identifying proteins, and the certainty of the results is computed by the number of masses matched for each protein [7]. Another useful tool is FindMod (http://www.expasy.ch/sprot/findmod.html) for the systematic characterisation of proteins using mass spectrometry [8].

We have created MS data forms that contain information used in mass spectrometry queries, summary information (Rank, MOWSE score, % Masses Matched, MW, *p*I, Species, Accession #, Protein Name) and additional information (Summary ID, Submitted Mass, Matched Mass, Delta

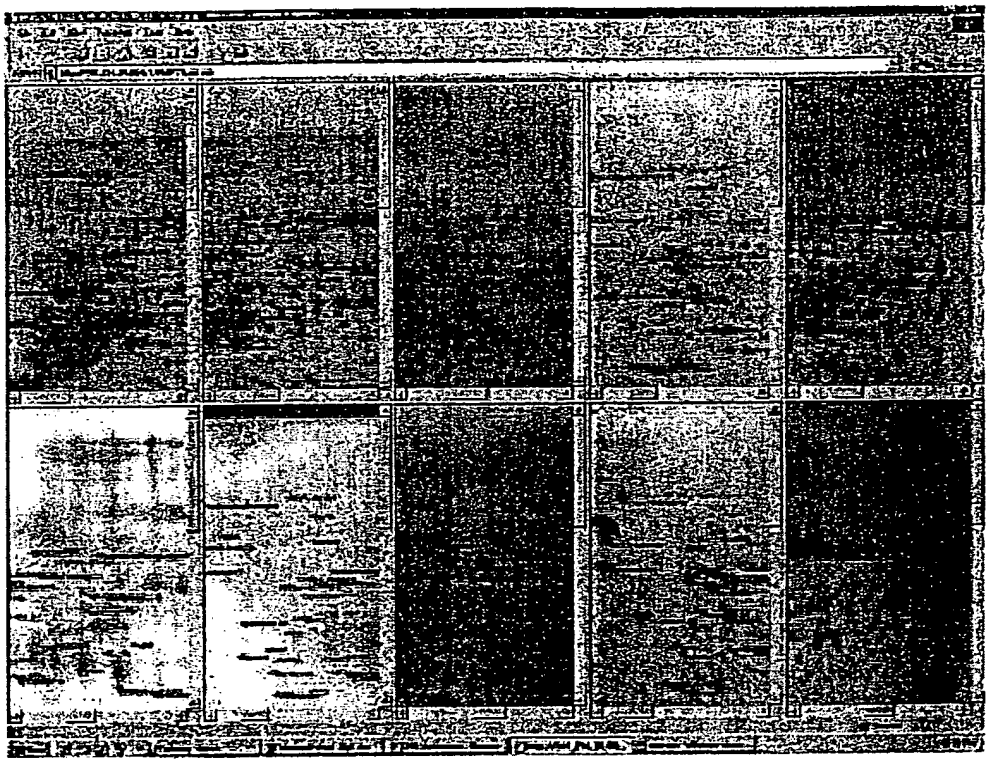


Figure 4. Virtual protein identification by clicking a spot.

PPM, Start, End, Peptide Seq, Modifications, Unmatched Masses). An example of the MS data form is shown in Fig. 5. Integrating the lung protein database with MS data provides a record of protein identification and high level of integration with other public databases, although substantial effort is required for data collection. We are currently evaluating an automated or semi-automated method of pulling these data when new information, which is relevant to our objectives, is available.

4.3 Integrating protein data with microarray data

As technology evolves, new computer aids and methods are introduced for genomic analysis as well as proteomic analysis. With respect to DNA microarray platforms, a current goal is to construct lung specific cDNA microarrays for lung cancer investigations. In the meantime RNA expression data for lung cancer is being collected using an Affymetrix oligonucleotide based system. This system automates the identification and quantification of microarray spots. Data files contain integrated intensities for each spot and ratios showing fold changes *per* spot. The use of oligonucleotide based microarrays for RNA analysis in lung cancer by our group has resulted in a massive amount of data. Integration of protein information in the lung protein database with microarray data allows us to extend data analysis capability to encompass RNA and protein data for a subset of genes.

5 Some findings derived from the lung cancer protein database

5.1 Unique proteomic pattern of small cell lung cancer

A major goal of our proteomic and genomic studies of lung cancer is to derive novel classification schemes that have utility in making a diagnosis, predicting outcome and in making therapeutic decisions. An important first step in this direction is to determine the ability of proteomic profiling to distinguish between known types of lung cancer. Specific protein differences between different types of cancer have been identified by other groups. In a recent study of breast, ovary and lung tumors, 20 differentially expressed proteins were identified [9] and in a prior study, 16 polypeptides were found to be associated with different histopathological features of lung cancer [10, 11]. In a study of 25 adenocarcinomas of the lung, 12 small cell

Spot #509
D:\Frank\MS\031299\0509-0.mis

Press stop on your browser if you wish to abort this MS-Fit search prematurely.

Sample ID (comment): Magic Bullet digest
Database searched: NCBInr 12.02.99
Molecular weight search (10000 - 45000 Da) selects 41693 entries
pI search (3.00 - 10.00) selects 213981 entries
Species search (HOMO SAPIENS) selects 22216 entries
Combined molecular weight, pI and species searches select 2950 entries
MS-Fit search selects 18 entries (results displayed for top 5 matches)
Considered modifications: | Peptide N-terminal Gln to pyroGlu | Oxidation of M | Protein N-terminus Acetylated |

Min. # Peptides to Match	Peptide Mass Tolerance (+/-)	Peptide Masses are	Digest Used	Max. # Missed Cleavages	Cysteines Modified by	Peptide N-terminus	Peptide C-terminus	Input # Peptide Masses
3	400.000 ppm	average	Trypsin	1	acrylamide(C)	Hydrogen	Free Acid (OH)	10

Result Summary

Rank	MOWSE Score	# (%) Masses Matched	Protein MW (Da)/pI	Species	NCBInr 12.02.99 Accession #	Protein Name
1	4.01e+005	9/10 (90%)	36638.7 / 5.71	HOMO SAPIENS	4557032 4557032	(X13794) lactate dehydrogenase B
2	291	3/10 (30%)	39152.5 / 8.76	HOMO SAPIENS	4505079 4505079	(U93163) MAGE-B1
3	238	3/10 (30%)	42695.9 / 9.55	HOMO SAPIENS	4503455 4503455	(AF022137) G protein-coupled receptor
4	241	3/10 (30%)	33021.9 / 8.27	HOMO SAPIENS	2135185 2135185	(S62286) fld-2-beta-I
5	215	3/10 (30%)	36915.2 / 8.90	HOMO SAPIENS	4585095 4585095	(L13689) putative

Detailed Results

1. 9/10 matches (90%), 36638.7 Da, pI = 5.71, Acc. # 4557032, HOMO SAPIENS, (X13794) lactate dehydrogenase B ;

m/z submitted	MH+ matched	Delta ppm	start	end	Peptide Sequence (Click for Fragment Ions)	Modifications
914.1086	914.1404	-34.7779	92	100	(K) IVVVTAGVRQKEGESRLNLVQRNVNVFKFIIPQIVKYSPDCIILVVSNPVDILTYVAWKISGFPKNRVIGSGCNLDSARFRYLMGERLGVHPLSCHGWILGEHGDSSVPVWSGVNVAGVSLKNLNPEMGTDKDPENWKEVHKQVVESAYEVIKLKGYTSWAIGLSVADLAESIMKNLRRVHPISTMIKGLYGIKEDVFLSVPCILNARGITSVINQKLKDDEVAQLRKSADTLWDIQKDLKDL(Q)	
953.3203	953.2367	87.2402	120	127	(K) FIPQIVKYSPDCIILVVSNPVDILTYVAWK(I)	
1638.8493	1638.8387	6.5058	44	58	(K) SLADELALVDVLEDKLKAELLALVDVLEDK(L)	
1695.9197	1695.9167	237.6337	8	23	(K) LIAPVAEEEATVPNNKITVVGVGAVGMASAISILGKSLADELALVDVLEDK(L)	
1988.7001	1988.4719	114.7577	24	43	(K) ITVVGVGAVGMASAISILGKSLADELALVDVLEDKLK(A)	
2004.8032	2004.4712	165.6136	24	43	(K)ITVVGVGAVGMASAISILGKSLADELALVDVLEDKLK(A)	Deamidation
2311.8481	2311.7372	47.9749	280	299	(K) GMYGIENEVFLSLPCILNARGMYGIENEVFLSLPCILNARG(G)	
2327.9638	2327.7365	106.2322	280	299	(K)GMYGIENEVFLSLPCILNARG(G)	Deamidation

Figure 5. MS data form

lung cancers, and 16 squamous cell tumors, by our group (manuscript submitted) an initial analysis of protein 2-D patterns uncovered a group of 52 protein spots that differed in average integrated intensity between the three groups. Performing simple two-sample *t*-tests gave *p* values of less than 0.05 for the 52 spots for at least one of the pairs of groups. Most of the spots differed between small cell and the remaining two diagnostic groups, with 47 spots differing significantly between small cell and adenocarcinoma groups and 44 between small cell and squamous ($p<0.05$). Between the adenocarcinoma and squamous groups 12 spots with difference of this significance were found. Summary data for some of the spots is presented in Table 4. The first two principal components of the data are graphed in Figure 6, and show that as a group the spots distinguish small cell tumors from the other two tumor types fairly easily.

We have identified 39 of this set of 52 spots by either *N*-terminal sequencing and/or MS of spot digests. Small cell lung cancers were characterized by higher average amounts for some proteins associated with cell prolifera-

Table 4. 39 identified protein spots found to differ between small cell, adenocarcinoma, and squamous tumors of the lung (n = 12, 25, 16). In the t-test columns are p values from the two–sided two–sample t-test comparing each pair of groups

Spot #	Unigene description	Offical gene symbol	Mean adeno-car-cinoma	Mean squa-mous	Mean small cell	t-test adenocar-cinoma *vs* small cell	t-test small cell *vs* squa-mous	t-test adenocar-cinoma *vs* squamous
294	vimentin	VIM	1.36	1.16	0.53	0.010	0.016	0.509
319	albumin	ALB	2.13	1.67	0.73	0.001	0.005	0.231
666	albumin	ALB	0.72	0.59	0.20	0.002	0.030	0.461
800	albumin	ALB	2.34	1.80	0.63	0.010	0.034	0.383
873	lectin, galactoside-binding, soluble, 1 (galectin 1)	LGALS1	1.95	1.69	0.83	0.000	0.002	0.310
928	ADP-ribosylation factor 1	ARF1	0.22	0.19	0.06	0.012	0.046	0.607
522	annexin A5	ANXA5	0.46	0.26	0.39	0.429	0.202	0.012
515	proliferating cell nuclear antigen	PCNA	0.15	0.18	0.36	0.002	0.011	0.464
577	stratifin	SFN	0.78	1.39	0.41	0.129	0.002	0.029
626	heat shock 27 kD protein1	HSPB1	0.87	1.18	0.30	0.000	0.002	0.128
631	heat shock 27 kD protein1	HSPB1	1.04	1.35	0.46	0.003	0.017	0.277
793	non-metastatic cells 1, protein (NM23A)	NME1	0.36	0.43	0.59	0.003	0.033	0.253
807	leukemia-associated phosphoprotein p18 (stathmin)	LAP 18	0.03	0.05	0.92	0.000	0.000	0.351
809	leukemia-associated phosphoprotein p18 (stathmin)	LAP18	0.55	0.50	3.88	0.000	0.000	0.732
931	S100 calcium-binding protein A9 (calgranulin B)	S100A9	0.95	1.18	0.24	0.026	0.001	0.447
104	protein phosphatase 2 (formerly 2A), regulatory subunit A (PR 65), beta isoform	PPP2R1B	0.17	0.13	0.65	0.000	0.001	0.188
110	procollagen-proline, 2oxoglutarate 4-dioxy-genase (proline 4-hydroxylase) beta polypeptide (protein disulfide isomerase; thyroid hormone binding protein p55)	P4HB	0.10	0.10	0.30	0.014	0.049	0.906
183	internexin neuronal intermediate filament protein, alpha	INA	0.04	0.04	0.16	0.000	0.000	0.751
229	tubulin, beta polypeptide	TUBB	0.14	0.27	0.83	0.000	0.000	0.028
289	keratin 15	KRT15	0.36	0.29	0.65	0.028	0.009	0.343
295	enolase 2, (gamma, neuronal)	ENO2	0.10	0.23	0.39	0.000	0.065	0.026
376	SET translocation (myeloid leukemia-associated)	SET	0.25	0.17	0.71	0.000	0.000	0.031
439	creatine kinase, brain	CKB	0.11	0.05	0.16	0.033	0.000	0.004
460	annexin A1	ANXA1	0.43	0.42	0.59	0.014	0.026	0.691
476	small glutamine-rich tetratricopeptide repeat (TPR)-containing	SGT	0.16	0.19	0.33	0.000	0.000	0.241
576	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase acctivation protein, epsilon polypeptide	YWHAE	0.40	0.38	0.82	0.000	0.001	0.697
579	tyrosine 3-monooxygenase/trypthophan 5-monooxgenase activation protein, theta polypeptide	YWHAQ	0.52	0.55	0.91	0.000	0.006	0.703
615	tyrosine 3-monooxygenase/ tryptophan 5-monooxygenase activation protein, zeta polypeptide	YWHAZ	0.93	1.09	1.79	0.000	0.003	0.336

Table 4. Continued

Spot #	Unigene description	Offical gene symbol	Mean adeno-car-cinoma	Mean squa-mous	Mean small cell	*t*-test adenocar-cinoma *vs* small cell	*t*-test small cell *vs* squa-mous	*t*-test adenocar-cinoma *vs* squamous
656	ubiquitin carboxyl-terminal esterase L1 (ubiquitin thiolesterase)	UCHL1	0.17	0.32	0.85	0.000	0.005	0.153
855	retinol-binding protein 1, cellular	RBP1	0.42	0.41	0.77	0.006	0.014	0.961
856	cellular retinoic acid-binding protein 2	CRABP2	0.25	0.38	0.63	0.000	0.017	0.037
902	enhancer of rudimentary (Drosophila) homolog	ERH	0.38	0.35	0.76	0.000	0.000	0.455
910	S100 calcium-binding protein A8 (calgarnulin A)	S100A8	1.46	1.43	0.35	0.040	0.001	0.950
934	keratin 17	KRT17	0.16	0.30	0.15	0.768	0.073	0.013
693	albumin	ALB	2.63	1.98	0.92	0.000	0.008	0.138
737	superoxide dismutase 2, mitochondrial	SOD2	1.17	1.22	0.54	0.013	0.001	0.836
789	collagen, type XV, alpha 1	COL15A1	0.57	0.50	0.26	0.031	0.186	0.658
906	S100 calcium-binding protein A11 (calgizzarin)	S100A11	2.95	2.62	0.53	0.000	0.000	0.506
924	lymphocyte cytosolic protein 1 (L-plastin)	LCP1	0.18	0.13	0.05	0.000	0.004	0.034

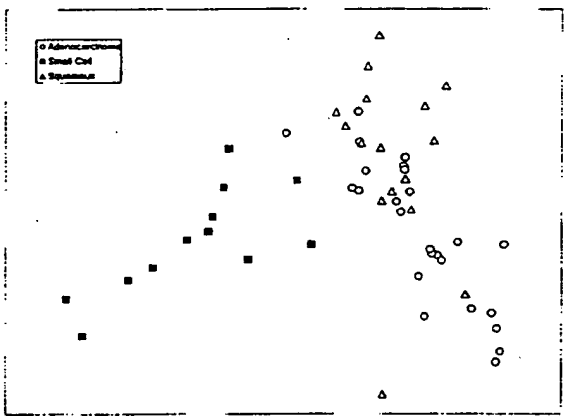


Figure 6. First two principle components for 52 protein spots distinguishing between lung tumor types. Small cell lung cancer samples are shown as squares, adenocarcinomas as circles and squamous lung tumors as triangles.

tion such as proliferating cell nuclear antigen (PCNA) and oncoprotein 18 (Op18) [12–15], particularly the once-phosphorylated form of Op18, as well as protein products of the UCHL1, RBP1, CRABP2, KRT15, and TUBB genes among others. Squamous cell and adenocarcinoma samples had greater amounts of the S100 proteins S100A8, S100A9, and S100A11, as well as larger average amounts of both the unphosphorylated and phosphorylated 27 kD heat shock protein (HSPB1). These two groups also had larger amounts of several protein spots detected on these gels that did not occur in similar gels made from cell lines and were thought to be cleavage products from proteins present in cells or plasma surrounding the tumor cells (*e.g.* cleaved albumin). The number of protein spots that differed between lung adenocarcinomas and squamous tumors were fewer than the number of proteins that distinguished between small cell lung cancer and the other two lung cancer types. ENO2 was smallest in the adenocarcinoma group, while ANXA5 and CKB were lowest and KRT17 and SFN highest in the squamous carcinoma samples. Several interesting spots found in the study remain to be definitively identified.

5.2 Correlations between RNA and protein expression

The availability of mRNA expression data from microarrays or Affymetrix chips for the same samples for which we have protein 2-D gel data permits several additional types of questions to be asked. We have thus far entertained only simple models of protein/mRNA relationships that ask which mRNA levels are most correlated with protein spot sizes. Figure 7 depicts such a correlation matrix using colors rather than numerical data, since this makes it easier to visualize the relationships. In cases for which the identity of the protein spot is known such investigations can answer the question of how well mRNA levels for a protein predict that protein's abundance. In cases of protein spots that have not yet been identified, or iden-

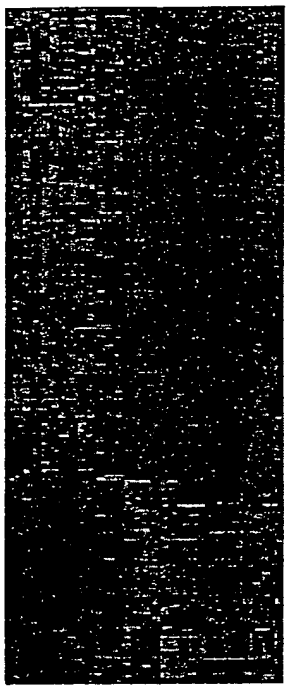


Figure 7. Correlation matrix of 30 protein spots (columns) with mRNA levels as measured by 200 probe-sets on Affymetrix HuFL chips. The correlation coefficients are depicted with colors, bright red being near-perfect correlation (r = 1) and bright green anticorrelation (r = –1). The figure was made using the TreeView software (rana.lbl.gov/EisenSoftware.htm).

tified without high confidence, such correlations can lead to or confirm hypothetical spot identifications. More generally one can search for larger groups of proteins and mRNA whose abundances are controlled by some common mechanism.

5.3 Identification of novel lung cancer markers

We have utilized a proteomic approach to identify proteins that commonly induce an antibody response in lung cancer. Such identified proteins or their corresponding autoantibodies likely have substantial utility for cancer diagnosis. There is also evidence that autoantibodies may be present prior to clinical diagnosis and therefore detection of autoantibodies or of circulating antigens may have utility for screening and early diagnosis of cancer. We have identified a battery of proteins that induce autoantibodies that are specific for different types of cancer. We have identified a panel of autoantibodies that are detectable in serum of lung cancer patients at the time of diagnosis. The availability of a database of protein expression in lung cancer has facilitated the identification of proteins that induce autoantibodies in addition to providing valuable information regarding the expression pattern of such antigens in different tumor types and cell lines. One such antigen we have identified in lung cancer is protein PGP 9.5 (Fig. 8) (Brichory *et al*, manuscript submitted) [16]. PGP 9.5 was identified as a protein in lung cancer that induces autoantibodies as part of a study in which sera from 64 newly diagnosed patients with lung cancer, from 99 patients with other types of cancer and from 71 noncancer controls were analyzed for antibody-based reactivity against lung adenocarcinoma proteins resolved by 2-D PAGE. Gels containing separated proteins were blotted and subsequently hybridized with individual sera from patients or controls. Unlike controls, autoantibodies against a protein identified by MS as protein gene product 9.5 (PGP 9.5) were detected in sera in 9 out of 64 patients with lung cancer.

Circulating PGP 9.5 antigen was detected in sera from two additional patients with lung cancer, without detectable PGP 9.5 autoantibodies. PGP 9.5 is a neurospecific polypeptide previously proposed as a marker for nonsmall cell lung cancer, based on its expression in tumor tissue. Using A549 lung adenocarcinoma cell line, we have demonstrated that PGP 9.5 was present at the cell surface, as well as secreted. Thus, the findings of PGP 9.5 antigen and/or antibodies in serum of patients with lung cancer suggest that PGP 9.5 may have utility in lung cancer screening and diagnosis, as part of a panel of such proteins or their corresponding antibodies, which we have identified.

6 Web pages

The relational database for storage of sample, image, protein information and other related data is being constructed in a stepwise fashion. The construction of a comprehensive database to collect all pertinent information is rather challenging and necessitates substantial resources. Similar effort in this area includes WebGel that is a web based gel database analysis system that contains previously quantified gel data generated from a stand-alone quantitative gel analysis system [16]. Public WebGel demonstration databases currently available can be found in the web site (http://www-lecb.ncifcrf.gov/webgel WebGel database). The task of web based retrieval of data from the protein database is rather complex as there are different kinds of data that may need to be retrieved. The microarray data could be stored in the database instead of Excel files, and the Access 2000 database that the MS team utilizes could be transferred to the database. Tables are being built to eliminate any handwritten collection of data. Developing a database is

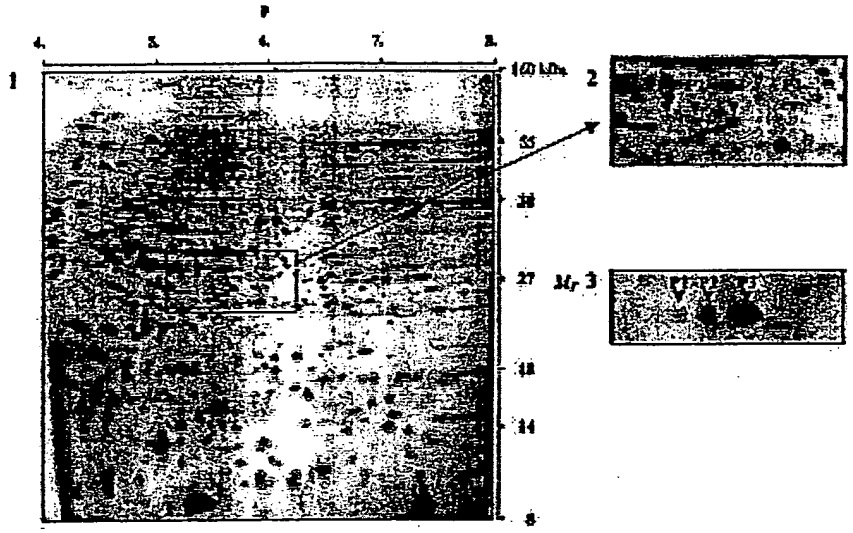


Figure 8. 2-D PAGE and Western blot analysis of A549 lung adenocarcinoma cell proteins. Panel 1 shows A549 2-D protein pattern after silver staining. The boxed area is shown in panel 2, in which arrows point to the location of PGP 9.5 forms (spots P1 to P3) recognized by sera from patients with lung cancer and the position of the form P4 recognized by a polyclonal rabbit anti-PGP 9.5 antiserum, which also recognizes P1-P3. Panel 2 shows close-ups of western blots hybridized with two different sera from patients with lung adenocarcinoma that showed reactivity against PGP 9.5 proteins.

hard because of complex and very large amount of unstructured data generated. There are conflicting pressures between "using what we've got already" and constructing something better. Sometimes there is a natural break in the data, such as when a shift is made from one platform type to another. Then one could "pile up" old data and organize it neatly. On the other hand, when new technologies are introduced, they require new ways of storing the data. The lung protein database is continuously evolving to enhance the relational schema to be more flexible and comprehensive and to make data processing more robust and automatic.

The lung protein database is a backbone to record proteome data for many different studies and to mine the existing data for new discoveries. The new generation LIPS provides investigators web-enabled interfaces to the laboratory databases and 2-D images with internet access. There is certainly a need for sharing information in the database on a global basis. We have used internet and WWW technologies to provide a distributed process with easy-to-use front-end user interface. Figure 9 shows a top level view of a web-based process for performing our studies from a data processing perspective. Some of our web pages were developed in Visual InterDev and ASP development environment on Microsoft and some were developed in Oracle 8i and WebDB web application environment on Solaris. As an example, the MS data web page is shown in Fig. 10. Detailed "how-to" documentation is provided as on-line help for recently extended capabilities of LIPS.

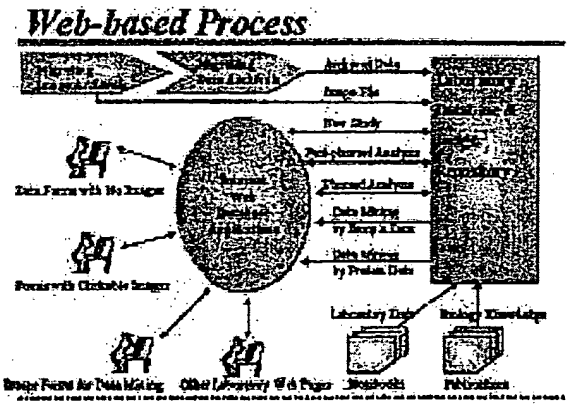


Figure 9. Web-based process of using lung protein database.

7 Conclusion

The value of the database we have constructed depends to a large measure on its content, the quality of data and the ease with which data can be retrieved and analyzed. While the amount of data generated is already quite sizeable, it is likely that the database will continue to undergo substantial expansion. Proteins are being identified at a rapid pace, thus enhancing our ability to link protein expression data with RNA based expression data for corresponding genes. As such, the database will play an important role in achieving our objective of developing novel classification schemes for lung cancer and the identification of novel markers for early diagnosis. The

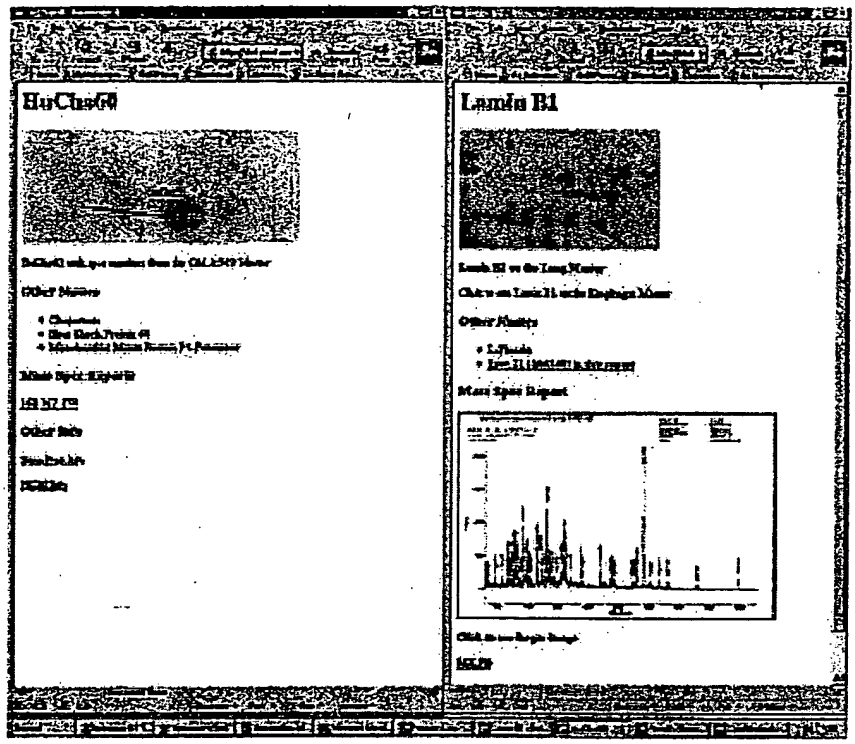


Figure 10. MS data web page.

database will also serve as a useful resource for other investigations of lung biology and of diseases other than lung cancer.

Received May 20, 2001

8 References

[1] Lemkin, P. F., *Electrophoresis* 1997, *18*, 2759–2773.
[2] Ali, I., Chan, Y., Kuick, R., Teichroew, D., Hanash, S. M., *Electrophoresis* 1991, *12*, 747–761.
[3] Oh, J. M. C., Hanash, S. M., Teichroew, D., *Electrophoresis* 1999, *20*, 766–774.
[4] Gras, R., Muller, M., Gasteiger, E., Gay, S., *et al.*, *Electrophoresis* 1999, *20*, 3535–3550.
[5] Fenyo, D., Qin, J., Chait, B. T., *Electrophoresis* 1998, *19* 998–1005.
[6] Fenyo, D., *Curr. Opin. Biotechnol* 2000, *11*, 391–395.
[7] Eriksson, J., Chait, B. T., Fenyo, D., *Anal. Chem.* 2000, *72*, 999–1005.
[8] Wilkins, M. R., Gasteiger, E., Gooley, A. A., Herbert, B. R., *et al.*, *J. Mol. Biol.* 1999, *289*, 645–657.
[9] Bergman, A. C., Benjamin, T., Alaiya, A., Waltham, M., *et al.*, *Electrophoresis* 2000, *21*, 679–686.
[10] Hirano, T., Franzen, B., Uryu, K., Okuzawa, K., *et al.*, *Br. J. Cancer* 1995, *72*, 840–848.
[11] Schmid, H. R., Schmitter, D., Blum, P., Miller, M., Vonderschmitt, D., *Electrophoresis* 1995, *16*, 1961–1968.
[12] Wang, Y. K., Liao, P.-C., Allison, J., Gage, D. A., *et al.*, *J. Biol. Chem.* 1993, *268*, 14269–14277.
[13] Melhem, R. F., Zhu, X. X., Hailat, N., Strahler, J., Hanash, S. M., *J. Biol. Chem.* 1991, *266*, 17747–17753.
[14] Zhu, X. X., Kozarsky, K., Strahler, J. R., Eckerskorn, C., *et al.*, *J. Biol. Chem.* 1989, *264*, 14556–14560.
[15] Hanash, S. M., Strahler, J. R., Kuick, R., Chu, E. H. Y., Nichols, D., *J. Biol. Chem.* 1988, *263*, 12813–12815.
[16] Lemkin, P. F., Myrick, J. M., Lakshmanan, Y., Shue, M. J., *et al.*, *Electrophoresis.* 1999, *20*, 3492–3507.